II AUTOMATIC INDEXING PROCEDURES

II . 1 Semantics and syntax

I shall consider automatic indexing under the four heads listed in Section I, namely analysis, description, index language generation, and searching, while recognising that operations fulfilling the same logical function may be carried out at different points in particular systems.

It is convenient to separate semantics and syntax, since in many systems they are clearly or effectively distinguished, even if ultimately there is little justification for labelling some conceptual category or relation syntactic or semantic.

The semantics of a retrieval system cover

1 from the document or request processing point of view

- a) in analysis, the identification of words or simple concepts
- b) in description, the provision of index terms or descriptors
- c) in searching, the exploitation of paradigmatic or class relations between terms;

2 from the index language point of view

- a) in generating the vocabulary, the choice of terms or descriptors
- b) in organising the vocabulary, the establishment of paradigmatic relations between terms.

The syntax of a retrieval system concerns

1 from the document point of view

- a) in analysis, the identification of syntactic roles and relations
- b) in description, the provision of relational information
- c) in searching, the exploitation of syntagmatic relations between terms;

2 from the language point of view

- a) in generating indicators, the choice of syntagmatic or logical relators to connect terms or descriptors
- b) in organising indicators, the establishment of modulation rules permitting changes to forms of syntactic structure.

As noted, the operations under most of these headings may be carried out partly or wholly automatically, and hence are covered by the label automatic indexing.

For example, for semantics, under

1a words may be extracted automatically using frequency criteria
b terms from a thesaurus may be assigned via the text density of entry words
c keywords may be connected through their statistical class relations;

2a descriptors may be provided by keyword grouping
b formal hierarchial relations may be established between term classes.

For syntax, under

1a nominal or verbal word groups may be identified by parsing
b logical relations may be associated with specific text syntax structures
c syntagmatic relations between terms may generalised;

2a}
b} automation dubious.

II.1.1 Documents and requests

In many cases, indexing procedures applied to documents may also be applied to requests (or profiles). For example, if a controlled indexing language is used, both documents and requests must be translated into it. In other cases, the direct treatment of documents and requests is not the same. For example, if documents are represented by simple lists of keywords, a Boolean request explicitly defines term relationships which are only implicit in the document descriptions. Another example is the use of truncated words in the search specification where the documents are indexed by full words. Although in a given search the effect is the same as if a matching document is indexed by truncated words, greater flexibility of indexing in relation to different searches is implied: truncating request words only allows alternative specifications with varying selective power. For example, if we search on 'comput-' we retrieve documents indexed by 'computer', 'computers', 'computing' and 'computation', but if we search on 'computer-' we retrieve the first two only.

We must therefore allow documents and requests to differ both in syntactic structure and in indexing vocabulary, though clearly differences of treatment cannot be carried too far. However, as most forms of description may be used for either documents or requests, references to document description should be generally interpreted to cover request description as well. Some analysis procedures, specifically statistical ones, are obviously not appropriate to requests, but these cases should be obvious. Specific remarks about requests will be included in the discussions of searching.

II . 2 Secondary system factors

There are many system factors affecting performance in general, and hence automatic indexing performance, but only indirectly; they are outside the actual indexing procedure. These factors may be assigned to two categories, less and more immediate.

II.2.1 Less immediate factors

These include gross system properties and constraints which may not be open to modification in the interests of indexing performance. For example,

1 re the collection:
its size (particularly in operational systems), subject matter, type of document, language etc., which may present particular problems if, for example, different types of document are included, or material in several languages;

2 re the mode of operation:
provision for different types of user and service from the same basic document material;

3 re the cost of the operation:
requirements to restrict costs, typically by adherence to simple techniques.

II.2.2 More immediate factors

These include factors affecting individual documents and requests.

Choice of input text

It is generally assumed that the choice of input text offered to the system for indexing affects performance. There are three obvious possibilities, full document text, abstract, and title. A number of comparisons to determine the relative value of these for manual indexing have been carried out, for example by Cleverdon 1966, Saracevic 1968, 1971 and Aitchison 1970. In general the results reported agree in finding titles as a base for indexing deficient in recall compared with abstracts or full text, though precision is good. Saracevic found a progressive increase in recall and decline in precision moving from titles to abstracts to text. Some comparisons with automatic indexing of a straightforward kind have also been carried out: Cleverdon and Aitchison compared titles and abstracts, and Salton 1968a titles, abstracts and full (short) text. The first two found titles inferior to abstracts on recall but superior on precision. Salton's rather different matching function showed an overall improvement in performance from titles to text; he claims there is less difference between abstract and text than between title and abstract.

Unfortunately different sources tend to be associated with different levels of indexing exhaustivity, so the comparisons are not confined to a single system factor. From an economic point of view it is of interest that though titles tend to perform indifferently, their defects may not be striking, and they are very cheap. If abstracts are machine held for independent reasons, for example for circulation in an SDI system, they may as well be used. It is not obvious how much is really gained from working with full text.

Degree of request vetting

The amount of work that is put into checking a request's intention before it is offered to the system may make a difference to the system's performance. In particular we can expect iterative searching with feedback to influence performance substantially. This is considered later.

Relevance

How much different interpretations of relevance affect system performance is not clear. Many experiments have allowed for different degrees of relevance, and control tests have been carried out with different relevance sets to check comparative indexing performance. On the whole we must expect grossly different relevance needs to suggest different approaches to indexing.

Matching functions

The matching function used during searching is logically independent of the form of indexing used; but its influence on performance may be considerable. Most retrieval systems use simple matching functions associated with Boolean request specifications, with the requirement that the full specification must be satisfied for a document to be retrieved. If terms are simply coordinated, and matches on decreasing subsets are allowed, we have the familiar coordination level approach. If terms are weighted, within this framework, we have notional coordination levels. However some workers have advocated more sophisticated matching functions leading to a more discriminating ranking of output, through normalisation to take account of varying document and request description length. This approach is adopted by the Smart project, and van Rijsbergen 1971, 1972. Salton 1968a,c argued that such functions improve performance, comparing performance for the overlap and cosine correlation functions for three collections: the cosine function performed noticeably better for two and no worse for the third. In some environments, where descriptions do not vary much, such functions would not be expected to perform better.